基于图遍历的局部社区发现算法 *

吴　建 [1]，王梓权 [1†]，易　亿 [1]，孙海霞 [2]

(1. 重庆邮电大学 通信与信息工程学院，重庆 400065；2. 西藏民族大学 信息工程学院，陕西 咸阳 712082)

摘　要：发现网络的社区结构对于了解复杂网络的结构和功能具有重要意义。针对当前局部社区发现算法扩张速度慢不适用于大规模网络的问题，提出了一种基于图遍历的局部社区发现算法。该算法首先找出网络中度数最低的节点，以该节点为起点通过影响力函数将网络中的节点分为社区节点和边界节点，形成初步的社区划分，然后通过适应度函数确定边界节点的社区得到最终划分结果。实验结果表明，该算法在真实网络上进行测试时不仅能够有效地挖掘网络中的社区结构而且具有较快的速度。

关键词：复杂网络；模块度；社区发现；图遍历

中图分类号：TB391　　doi: 10.3969/j.issn.1001-3695.2018.03.0157

Local community detection algorithm based on Graph Traversal

Wu Jian[1], Wang Ziquan[1†], Yi Yi[1], Sun Haixia[2]

(1. *College of Communication & Information Technology*, *Chongqing University of Posts & Telecommunications*, *Chongqing* 400065, *China*; 2. *College of Information Engineering*, *Xizang Minzu University*, *Xianyang Shaanxi* 712082, *China*)

Abstract: Detection of communitiy structure is significant in understanding the structures and functions of the complex network. In view of the problem of the slowness of the community diffusion and not suitable for large-scale network, The paper proposed a local community detection algorithm based on breadth first traversal. The algorithm finds out the node with lowest degree in the network, and uses this node as a starting point to divide the nodes into community nodes and boundary nodes to form the initial community detection by influence function. Then use fitness function to get the final cover. The experimental results show that the algorithm tested in a real network can effectively dig out community structure in the network and have faster speed.

Key words: complex network; modularity; community detection; graph traversal

0　引言

社区结构是复杂网络无处不在的拓扑性质之一，寻找复杂网络的社区结构已经成为一个基本问题。社区结构在控制复杂网络、分析复杂网络拓扑、预测网络中个体行为以及深入理解网络功能等方面发挥着非常重要作用。这种结构通常表现为一个网络能被划分为许多组,其中组内成员的联系较为密切,而组与组之间成员的联系相对稀疏[1-3]。所谓社区发现是利用网络中包含的拓扑信息来发现网络中的模块化结构。发现网络中的社区结构对于控制疾病和网络病毒的传播具有极其重要意义。

因为科研人员对社区结构定义的理解不同，所以在进行社区划分时遵循的标准也各不相同，可以分为两大类：全局社区发现算法和局部社区发现算法 [4]。全局社区发现算法在进行社区划分前需要获取网络的全局信息，从全局角度出发进行社区划分，算法结果可信度高[5]。

随着大数据时代的到来，复杂网络的规模逐渐增大，网络的全局信息逐渐变得难以把握，全局社区发现算法已经难以适用于规模较大的网络，因此，科研人员提出了局部社区发现算法。局部社区发现算法无需事先获得网络的全局信息，相反，它是通过网络中节点（子图）的局部信息对社区进行扩展。局部社区发现算法无需事先获取整个网络的全局信息就能够进行社区划分。与全局社区算法相比，局部社区发现算法在运行时间、应用范围、算法灵活性等方面存在更大优势，已成为当前研究的热点。目前，根据划分思想不同，可以分为标签传播算法、派系过滤算法以及局部扩张算法。标签传播算法（LPA 算法）首先给网络中所有节点分配对应的标签，然后依据传播规则反复更新节点的标签，直至整个网络中节点的标签不再改变。最后，拥有同样标签的节点即为一个社区。由于 LPA 算法规定每个节点只能同时拥有一个标签，所以该算法无法挖掘出网络的重叠结构。针对这一问题，Steve Gregory 等人提出了 COPRA

收稿日期：2018-03-12；**修回日期：**2018-04-25　**基金项目：**国家自然科学基金资助项目（61571071）

作者简介：吴建（1970-），男，重庆人，正高级工程师，主要研究方向为计算机图像处理；王梓权（1992-），男（通信作者），山东淄博人，硕士研究生，主要研究方向为数据可视化、社会网络（1225073656@qq.com）；易亿（1993-），男，湖北黄冈人，硕士研究生，主要研究方向为深度学习、人工智能；孙海霞（1972-），女，甘肃秦安人，副教授，主要研究方向为计算机应用技术、网络技术等.

算法[6]，该算法允许节点同时拥有多个标签，即可以同时存在于多个社区中，进而发现网络中的重叠结构。派系过滤算法是基于渗透思想，该算法定义 k-clique 是网络中包含 k 个节点的全连通子图，而网络中的社区是具有共享节点的全连通子图的集合。该算法首先搜索网络中所有 k-clique 然后建立以 k-clique 为核心新图，在该图中如果两个 k-clique 拥有 k-1 个共有节点则为它们建立一条边，最终，每个连通子图即为一个社区。局部扩张是以给定的节点（子图）为中心，根据传播规则和判断函数逐步合并周围邻居节点，以发现社区结构。例如，Lancichinetti 等人提出的 LFM 算法[7]，该算法以种子节点为中心，通过度量函数来发现网络中的社区结构。由于 LFM 算法通过随机选取网络中的节点作为种子节点，结果稳定性较差，需要多次运行取最优结果。对于这一问题， Conrad Lee 等人提出了 GCE 算法[8]，该算法首先寻找网络中的 k-clique，再使用 LFM 算法的适应度函数对这些 k-clique 进行扩展。GCE 算法虽然稳定性高，但运行时间较长，划分结果的好坏很大程度上依赖于 k 值的选取。

目前，局部社区发现算法大多是将邻居节点逐个随机加入社团中，每次迭代只能够添加或删除一个节点，使得重复计算次数较多，效率较低，社团扩张速度较慢，难以适用于大规模网络。针对这一问题，本文提出了基于图遍历的局部社区发现算法，其基本思想是：把度数最小的节点作为起始节点，根据提出的影响力函数与设定的阈值 r 对网络中的节点进行标记，形成初步的社区划分，然后通过适应度函数 F 来的到最终结果。

1 基于图遍历的局部社区发现算法

由文献[6，11,12]可知，目前大多数的局部社区发现算法在进行社区划分时，每次只能判断一个节点的社区归属，为了发现网络中的重叠社团结构还需要对某些节点进行重复判断，这样会导致算法在进行社区划分时迭代次数过多，大大增加了算法的运行时间，难以适用于大规模网络和社团数目较多的网络。针对这一问题，本文提出了基于图遍历的局部社区发现算法，解决现有局部社区算法效率较低的问题。

1.1 基本定义

为了准确理解本文算法，首先给出如下定义.

定义 1 网络表示形式。假设 G 为一个网络表示为 G(V,E)，其中 $V=\{v_1,v_2,v_3,\ldots,v_n\}$ 代表节点的集合，$E=\{e_1, e_2, e_3, \cdots e_m\}$ 代表边的集合。

定义 2 节点v受到的影响力 NIS(v)为

$$NIS(v) = \frac{N(v)}{Degree(v)} \quad (1)$$

其中：N(v)表示 v 的邻居节点中受影响节点的数目；degree(v)表示节点 v 的度数。显然，NIS(v)的值始终在 0~1，对于起始节点，它的 NIS 值为是零。如果节点 v 的所有邻居节点都受到影响，那么 NIS(v)将是 1。

定义 3 边界节点。如果节点 v 受到的影响力 NIS(v)小于一个预定的阈值 r（$0 \le r \le 1$），那么这个节点被认为是一个边界节点(Border Node), 节点标记为 BN。

定义 4 社区节点。如果节点 v 受到的影响力 NIS(v)大于或等于一个预定的阈值 r（$0 \le r \le 1$），那么这个节点被认为是一个社区节点，节点的标记视具体情况而定。

定义 5 重叠模块度。重叠模块度（EQ）是 Shen 等人提出的一种基于模块度改进的函数[9]，该函数解决了普通模块度（Q）[10]无法评价重叠社区的问题。重叠模块度函数考虑了同时隶属于多个社区之间的重叠节点，由于此类节点同时隶属于复数社区，在计算模块度时应该削弱此类节点对社区结构紧密度的影响，基于节点的重叠度越高，它对模块度的影响就越小这一原则，定义了重叠模块度，公式如下：

$$EQ = \frac{1}{2m}\sum_{c}\sum_{i,j\in c}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\frac{1}{O_i O_j} \quad (2)$$

其中：A_{ij}代表网络中邻接矩阵的元素值；m代表网络边的总数；O_i代表节点i所属的社区个数；k_i表示节点i的度数。与Q函数相比，EQ函数弱化了重叠节点对社区结构紧密度的影响，当不存在重叠节点时，EQ函数与Q函数等价。

定义 7 适应度函数。适应度函数是由 Lancichinetti 提出用于衡量社区内外连接程度优劣的函数，具体公式如下：

$$F = \frac{K_{in}^G}{K_{in}^G + K_{out}^G} \quad (3)$$

其中：K_{in}^G 为社区G内部度数；K_{in}^G 为社区G外部度数。

1.2 初始社区划分

在初始社区划分过程中，该算法定义了一种新的节点标记规则，能够一次判断选中节点所有邻居节点所属社区，对网络进行初始社区划分，减少算法迭代次数，提高算法效率。经过初始社区划分后，网络中的节点被分为边界节点与社区节点 2 大类。其中，边界节点由于受到影响力小，暂时无法判断其所属社区，需要进一步判断，而拥有相同标签的社区节点则被认为已经稳定，在进行社区最终划分时只需判断边界节点，减小了后续阶段需判断的节点数目。

初始社区划分标记规则如下：

a)起始节点一定是边界节点，标记为 BN。

b)如果一个节点为边界节点且通过该节点发现其邻居节点为也为边界节点，邻居节点的标记为 BN。

c)如果一个节点为边界节点且通过该节点发现其邻居节点为社区节点，则将该节点从边界节点改为社区节点，其标记从 BN 更新该节点初始的标签，其邻居节点的标记与该节点的此时的标签相同。

d)如果一个节点为社区节点且通过该节点发现其邻居节点为边界节点，则将邻居节点的标记为 BN，该节点的标记保持不

变。

e)如果一个节点为社区节点且通过该节点发现其邻居节点也为社区节点，将其邻居节点的标记为该节点当前的标记。

f)每个节点只允许携带一个标记。

1.3　最终社区划分

经过初始划分后，网络中的节点被分为社区节点和边界节点两类，为了判断边界节点所属社区，发现网络中的重叠结构，本文采用 Lancichinetti 定义的适应度函数 F，将边界节点分别加入到其相邻的社区节点中，并计算合并前后的 F 值，若合并后的 F 值大于合并前的 F 值，那么就为边界节点添加一个合并社区的标记，每个边界节点可以拥有多个标记。当所有边界节点合并完毕后，去网络中掉所有的 BN 标记，此时拥有相同标记的节点被划分在同一社区。

1.4　算法步骤

输入：网络 G(V,E)，阈值 r

输出：网络 G 的社区划分结果

a)计算网络中所有节点的度数，选取度数最小的节点作为起始节点，若存在多个度数相同的节点则从中随机选择一个作为起始节点，由于此时网络中的节点都未被标记，因此该节点被标记为边界节点，标记为 BN;

b)由起始节点开始，根据式（1）计算该节点周围所有邻居节点受到的影响力，并根据节点标记规则对邻居节点进行的标记；

c)判断当前已标记节点是否还有未被标记的邻居节点，若存在，则根据式（1）计算所有未标记邻居节点受到的影响力根据规则给与相应的标记，再次执行步骤 c);若不存在，则跳转至步骤 d);

d)选择网络中的边界节点与邻居社区进行合并，根据式（3）计算每次合并前后的适应度函数（F），若合并后新社区的 F 大于合并前的社区 F，则按照规则对合并后的节点标记更新；

e)当网络所有边界节点均合并完毕时区掉所有的 BN 标记，将拥有相同标记的节点被划分到统一社区中。

1.5　时间复杂度分析

由上文可知，本文算法主要可以分为初始社区划分、最终社区划分两个阶段。对于一个包含 n 个节点和 m 条边的网络，寻找起始节点需要计算网络中所有节点的度数并将度数，找出度数最小的节点作为起始节点，时间复杂度为 O(n); 社区初始划分需要计算网络中所有节点的 NIM 值，并根据 NIM 值的大小与阈值 r 的大小对节点进行标记，这一阶段时间复杂度为 O(n); 最终社区划分需要将相同标签的社区节点压缩为一个个新节点并将网络中的节点进行合并，计算每次合并后的 F 值，直到 F 无法再增大，这里考虑最坏情况，即每个所有节点均为独立的社区需要两两合并，这种情况时间复杂度为 O(m)。最终，本文算法的时间复杂度为 O(n+m)。对于稀疏网络，本文算法的复杂度可近似为 O(n)。

2　算法测试

2.1　阈值分析

由算法思想可知本文算法中阈值 r 的大小对社区结构的优劣与社区最终的划分结果有重要影响，因此首先要确定阈值 r 的初始值，这里，代表性的真实数据。其中，真实数据分别是 Karate 数据集、Football 数据集、Dolphins 数据集以及 Netscience 数据集，其详细信息如表 1 所示。

表 1 中 Karate 数据集是通过对美国大学某俱乐部成员的社交情况进行观测而构建的社会网络，节点表示俱乐部的成员，节点间的边表示成员之间存在的友谊关系；Football 数据集是科学家 Newman 等人根据美国大学生足球联赛而构建的一个复杂社会网络，网络中的节点代表足球队，节点之间的边表示两只球队之间进行过比赛；Dolphin 数据集是 Lusseau 等人根据新西兰海豚群体的交流情况而得到的海豚社会关系网络，其中节点表示海豚，而边表示海豚间的频繁交流；Netscience 数据集是一个从事网络理论研究和实验的共同作者网络，其中，节点代表研究人员，两节点之间的边表示两位研究人员曾经合著过文章。

通过在 4 个数据集上分别运行 30 次求阈值 r 的平均值发现当 r 的范围在 0.4≤r≤0.8 时得到的 EQ 值较好，当阈值 r=0.7 时在大多数据集上得到的社区结构最佳（如图 1）。因此，在接下来的实验中本文选择阈值 r 的值为 0.7。

表 1　数据集信息

Number	Dataset	Nodes	Edges
1	Karate	34	78
2	Football	115	613
3	Dolphins	62	159
4	Netscience	1461	2742

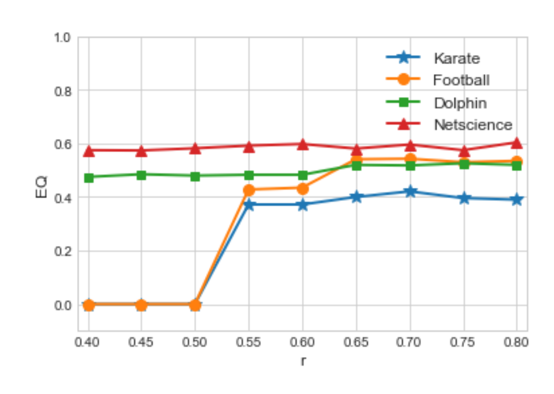


图 1　EQ 值与阈值 r 的关系

2.2　人工网络实验

为了更好的阐述算法过程，本文在构造的简单网络上进行了实验（图 2）。表 2 显示了在阈值 r=0.7 时本文算法在初始社区划分时不同阶段每个节点的标记。可以看出以节点 8 为起始节点经过初始社区划分后，网络被划分为边界节点和社区节点两大类。其中，边界节点有节点 3、 8、13，社区节点为拥有标

记 7 的节点 5、7，拥有标记 0 的节点 0、1、2、4，拥有标记 10 的节点 9、10、11、12。在进行最终划分时只需将无需再判断社区节点，只需将边界节点加入其邻居社区中根据公式（3）计算合并前后的 F 函数，判断其最终所属社区，该网络最终社区划分结果如表 3 所示。

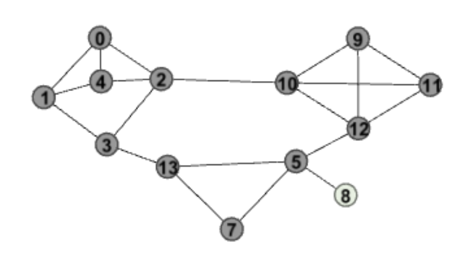


图 2 简单网络

表 2 简单网络初始社区划分

过程	节点名称	NIS	节点类型	标记
1	8	0	边界节点	BN
2	5	0.25	边界节点	BN
3	13	0.67	边界节点	BN
	7	1	社区节点	7
	12	0.25	边界节点	BN
	5		社区节点	7
4	3	0.33	边界节点	BN
	10	0.75	社区节点	10
	9	1	社区节点	10
	11	1	社区节点	10
	12		社区节点	10
5	1	0.33	边界节点	BN
	2	0.5	边界节点	BN
6	0		社区节点	0
	4	1	社区节点	0
	1	1	社区节点	0
	2		社区节点	0

表 3 简单网络最终社区划分

社区编号	节点编号
社区 1	5,8,13,7
社区 2	9,10,11,12
社区 3	0,1,2,3,4

2.3　大规模网络数据集实验

为了验证本文算法在大规模数据集上的有效性，将本文算法与 LFM、CPM 以及 COPRA 算法进行了比较，社区质量评价函数选取 EQ 函数作为评价标准，选取的真实网络数据信息如表 4 所示。

表 4　真实网络数据描述

Dataset	Nodes	Edges
blogs2	30.5K	82.3K
Enron	33.7K	181K
Amazon	335K	926K

表 5 列出了本文算法和其他三个算法在三个大规模数据集下的测试结果。实验结果表明，本文算法在不同的真实数据集上均能取得了相对较好的 EQ 值，划分社区需要的运行时间比其余 3 个算法要少。LFM 算法由于随机选择网络中的节点作为种子节点，在实验过程中，该算法每次运行的 EQ 值都不相同，容易受到种子节点位置的影响，且在进行社区划分时每次只能判断一个节点，随着网络规模增大其算法运行时间增长较为明显；CPM 算法在大规模数据集下能够挖掘出质量较高的社区，但是寻找网络中的 k-clique 需要较长的时间；COPRA 算法由于需要不定时的迭代更新网络中节点的标签耗时较多，在大规模网络下运行的效果并不理想；本文算法在初始社区划分阶段能够一次判断选中节点所有邻居节点所属社区减少算法迭代次数，形成初步划分，减少最终社区划分需要判断的节点数目。

表 5 真实网络划分结果

网络		本文算法	LFM	CPM	COPRA
blogs2	EQ	0.487	0.365	0.472	0.436
	运行时间	1.96s	11.8s	3.88s	81.14s
Enron	EQ	0.49	0.196	0.54	0.361
	运行时间	4.89s	16.4s	10.3s	96.41s
Amazon	EQ	0.547	0.642	0.534	——
	运行时间	4.85min	23min	6.63min	——

3　结束语

本文提出了一种基于图遍历的局部社区发现算法，该算法定义了影响力函数 NMI，根据节点受到的影响力与设定的阈值 r 来同时对多个节点的标签进行更新形成初始社区划分，然后通过适应度函数来得到最终的社区划分，减少了算法在判定节点所属社区时的重复迭代次数，提高了算法运行效率。实验证明，本文算法划分出的社区质量虽然受到阈值 r 的大小的影响，但总体上得到的 EQ 值与上述算法相比较高，即本文算法划分出的网络社区结构更加准确，并且拥有较低的时间复杂度。后续工作将考虑到网络中边的权值问题，使算法能够应用到加权网络当中。

参考文献：

[1] Zhang Xingyi, Wang Congtao, Su Yansen, *et al.* A fast overlapping community detection algorithm based on weak cliques for large-scale networks [J]. IEEE Trans on Computational Social Systems, 2017, 4 (4): 218 – 230.

[2] Vespignani A. Complex networks: The fragility of interdependency. [J]. Nature, 2010, 464 (7291): 984-5.

[3] 王琦，温志平. 一种基于多维遗传算法的重叠社区发现方法 [J]. 计算机应用研究，2016, 33 (12): 3543-3546. (Wang Qi, Wen Zhiping. Multidimensional genetic algorithm for overlapping community detection [J]. Application Research of Computers, 2016, 33 (12): 3543-3546.)

[4] 刘立寒，方志祥，萧世伦，等. 带源节点的快速社区发现算法 [J]. 计

算机工程与应用, 2016, 52 (23): 75-80. (Liu Lihan, Fang Zhixiang, Xiao Shilun, *et al.* Fast communities detection algorithm with source nodes [J]. Computer Engineering and Applications, 2016, 52 (23): 75-80.)

[5] 李建华，汪晓锋，吴鹏. 基于局部优化的社区发现方法研究现状 [J]. 中国科学院院刊, 2015 (2): 238-247. (Li Jianhua, Wang Xiaofeng, Wu Peng. Review on Community Detection Methods Based on Local Optimization [J]. Bulletin of Chinese Academy of Sciences, 2015 (2): 238-247.)

[6] Kumpula J M, Kivelä M, Kaski K, *et al.* Sequential algorithm for fast clique percolation [J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2008, 78 (2): 026109.

[7] Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure of complex networks [J]. New Journal of Physics, 2012, 11 (3): 19-44.

[8] Lee C, Reid F, McDaid A, *et al.* Detecting highly overlapping community structure by greedy clique expansion [J]. arXiv preprint arXiv: 10021827, 2010.

[9] Shen Huawei, Cheng Xueqi, Cai Kai, *et al.* Detect overlapping and hierarchical community structure in networks [J]. Physica A Statistical Mechanics & Its Applications, 2009, 388 (8): 1706-1712.

[10] Blondel V D, Guillaume J L, Lambiotte R, *et al.* Fast unfolding of communities in large networks [J]. Journal of Statistical Mechanics, 2008, 2008 (10): 155-168.

[11] Wang Xiaofeng, Liu Gongshen, Li Jianhua. Overlapping community detection based on structural centrality in complex networks [J]. IEEE Access, 2017, 5: 25258-25269.

[12] 齐金山，梁循，王怡. 基于种子节点选择的重叠社区发现算法 [J]. 计算机应用研究, 2017, 34 (12): 3534-3537. (Qi Jinshan, Liang Xun, Wang Yi. Overlapping community detection algorithm based on selection of seed nodes [J]. Application Research of Computers, 2017, 34 (12): 3534-3537)